Clustering Problems for More Useful Benchmarking of Optimization Algorithms

Marcus Gallagher*

School of Information Technology and Electrical Engineering,
The University of Queensland, Brisbane, 4072. Australia
marcusg@itee.uq.edu.au

Abstract. This paper analyses the data clustering problem from the continuous black-box optimization point of view and proposes methodological guidelines for a standard benchmark of clustering problem instances. Clustering problems have been used many times in the literature to evaluate evolutionary, metaheuristic and other global optimization algorithms. However much of this work has occurred independently and the various experimental methodologies used have produced results which tend to be incomparable and provide little collective wisdom as to the difficulty of the problems used, or an objective measure for comparing and evaluating the performance of algorithms. This paper surveys some of the clustering literature and results to identify issues relevant for benchmarking. A set of 27 problem instances ranging from 4-D to 40-D and based on three well-known datasets is identified. To establish some pilot results on this benchmark set, experiments are presented for the Covariance Matrix Adaptation-Evolution Strategy and several other standard algorithms. A web-repository has also been created for this problem set to facilitate better experimental evaluations of algorithms.

Keywords: Algorithm Benchmarking, Continuous Black-box Optimization, Clustering.

1 Introduction

In evolutionary computation and metaheuristic optimization, an enormous number of algorithms have been developed. Since no algorithm is superior in the theoretical, No Free Lunch sense, in practice the performance differences we observe depend on how well the mechanisms of the algorithm match the structure of the problem landscape. A key step towards understanding the matching between problems and algorithms is to develop better benchmark problems and more rigorous approaches to the experimental analysis of algorithms. Unfortunately, the dominating paradigm in the literature has been to continually develop new algorithm variants and to evaluate these techniques in isolation. For continuous

* The author acknowledges the contribution of the Dagstuhl Theory of Evolutionary Algorithms Seminar 13271 (http://www.dagstuhl.de/13271/) to the work in this paper.

G. Dick et al. (Eds.): SEAL 2014, LNCS 8886, pp. 131–142, 2014.

black-box optimization, artificial test functions (e.g. Sphere, Rastrigin, Rosenbrock) have been used hundereds of times, but a question such as “what is the best performance of a black-box optimization algorithm on function f, given 10^5 functions evaluations?” seems to be difficult (if not impossible) to answer using the literature. The situation is even more problematic, because subtle differences in experimental settings in different papers (e.g. using a different bound on the feasible search space) mean that results are often not strictly comparable. Recently, research has begun to focus more on such experimental issues. For example, the Black-Box Optimization Benchmarking (BBOB) problem set[1] resolves many of these issues by standardizing many aspects of the experimental setting. However, it is very important to also evaluate algorithms on real-world problems, since it is difficult to know how well artificial test problems represent real-world problems and hence to what extent algorithm performance on artificial problems is indicative of real-world performance. It can be difficult to use real-world problems for algorithm benchmarking because real problems may require expert domain knowledge to configure, or may come with additional complexities that are not part of the basic optimization algorithm (e.g. complex constraints). Ideally, problems that are real-world “representative” while being convenient for benchmarking should provide a valuable contribution to experimental research practice.

This paper examines data clustering as a useful source of continuous, black-box benchmark problems. In Section 2, the sum of squares clustering problem is defined and its key properties discussed. Clustering problems have previously been used in the literature to test optimization algorithms: Section 4 reviews some of this literature and discusses why it is difficult to compare with previously reported results. A specification is proposed to describe clustering problem instances and a set of problem instances defined (and made available via the web). To establish some baseline results for future comparison, a number of commonly-used algorithms are applied to the clustering problem sets. The experiments are described in Section 5 and Results presented in Section 6. Where possible, the results are also compared with previous results from the literature, revealing some surprising insights. The work is summarised in Section 7.

2 Clustering

The sum of squares clustering problem (see, e.g.[13]) can be stated as follows. Given a set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq I\!R^d$ of n data points, find a set of k cluster centers $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_k\} \in I\!R^d$ to minimize:

$$f(\mathcal{C}|\mathcal{X}) = \sum_{i=1}^{n} \sum_{j=1}^{k} b_{i,j} ||\mathbf{x}_i - \mathbf{c}_j||^2$$

where

$$b_{i,j} = \begin{cases} 1 & \text{if } ||\mathbf{x}_i - \mathbf{c}_j|| = \min_j ||\mathbf{x}_i - \mathbf{c}_j|| \\ 0 & \text{otherwise.} \end{cases}$$

[1] http://coco.gforge.inria.fr/doku.php